SCHOOL OF
COMPUTING

Title: A Systematic Evaluation of Evidence-Based Methods in Cyber Security User Studies

Names: Kovila Coopamootoo and Thomas Gross

TECHNICAL REPORT SERIES

No. CS-TR-1528 **July 2019**

TECHNICAL REPORT SERIES

No. CS-TR-1528 July 2019

Title: A Systematic Evaluation of Evidence-Based Methods in Cyber Security User Studies

Authors: Kovila Coopamootoo and Thomas Gross

Abstract: Background. In the recent years, there has been a movement to strengthen evidence-based methods in cyber security under the flag of "science of security." It is therefore an opportune time to take stock of the state-of-play of the field. Aim. We evaluated the state-of-play of evidence-based methods in cyber security user studies.
Method. We conducted a systematic literature review study [1] of cyber security user studies from relevant venues in the years 2006–2016. We established a qualitative coding of the included sample papers with an a priori codebook of 9 indicators of reporting completeness [2]. We further extracted effect sizes for papers with parametric tests on differences between means for a quantitative analysis of effect size distribution and post-hoc power. Results. We observed that only 21% of studies replicated existing methods while 78% provided the documentation to enable future replication. With respect to internal validity, we found that only 24% provided operationalization of research questions and hypotheses. We observed that reporting did largely not adhere to APA guidelines as relevant reporting standard [3]: only 6% provided comprehensive reporting of results that would support meta-analysis. We, further, noticed a considerable reliance on p-value significance, where only 1% of the studies provided effect size estimates [4]. Of the tests selected for quantitative analysis, 80% reported a trivial to small effect, while only 28% had post-hoc power ($1 - \beta \geq 80\%$). Only

16% were still statistically significant after Bonferroni correction for the multiple-comparisons made.
Conclusions. This study offers a first evidence-based reflection on the state-of-play in the field and indicates areas that could help advancing the field's research methodology.

Printed and published by Newcastle University,
School of Computing, Urban Sciences Building,
Newcastle upon Tyne, NE4 5TG, England.

Bibliographical Details :

Title and Authors

A Systematic Evaluation of Evidence-Based Methods in Cyber Security User Studies

Kovila Coopamootoo and Thomas Gross

NEWCASTLE UNIVERSITY
School of Computing. Technical Report Series. CS-TR-1528

Abstract

Background. In the recent years, there has been a movement to strengthen evidence-based methods in cyber security under the flag of “science of security.” It is therefore an opportune time to take stock of the state-of-play of the field. Aim. We evaluated the state-of-play of evidence-based methods in cyber security user studies.

Method. We conducted a systematic literature review study [1] of cyber security user studies from relevant venues in the years 2006–2016. We established a qualitative coding of the included sample papers with an a priori codebook of 9 indicators of reporting completeness [2]. We further extracted effect sizes for papers with parametric tests on differences between means for a quantitative analysis of effect size distribution and post-hoc power.

Results. We observed that only 21% of studies replicated existing methods while 78% provided the documentation to enable future replication. With respect to internal validity, we found that only 24% provided operationalization of research questions and hypotheses. We observed that reporting did largely not adhere to APA guidelines as relevant reporting standard [3]: only 6% provided comprehensive reporting of results that would support meta-analysis. We, further, noticed a considerable reliance on p-value significance, where only 1% of the studies provided effect size estimates [4]. Of the tests selected for quantitative analysis, 80% reported a trivial to small effect, while only 28% had post-hoc power ($1 - \beta \geq 80\%$). Only 16% were still statistically significant after Bonferroni correction for the multiple-comparisons made.

Conclusions. This study offers a first evidence-based reflection on the state-of-play in the field and indicates areas that could help advancing the field’s research methodology.

About the authors

Kovila P.L. Coopamootoo is a Research Fellow at Newcastle University's Secure Resilient Systems (SRS) group. Her research topic addresses evidence-based research for human factors of cyber security and privacy. Prior to her fellowship, she was a Lecturer in Cyber Security at the University of Derby for a short while. Prior to that, she was a Post-Doctoral Research Associate at Newcastle (employed on EPSRC Hyper-DoVe, EU FutureID and

EU PrismaCloud projects), Co-I on the Newcastle Human Dimensions in Cyber Security Lab and Co-I on an International Research Collaboration Award. She has designed and reported on a number of evidence-based studies and has supervised and co-supervised a number of students at undergraduate, MSc and PhD levels.

Thomas Gross is a Reader in System Security. He is Director of the Centre for Cybercrime and Computer Security (CCCS) and the Newcastle Academic Centre of Excellence in Cyber Security Research (ACE-CSR). He is the Principal Investigator of the ERC Starting Grant CASCAde (Confidentiality-Preserving Security Assurance, GA no 716980). He is the Principal Investigator and Director of the EPSRC Contrails Centre CRITiCaL (Northern Cloud Crime Centre). He has published over 50 cybersecurity articles, has 12 granted patents and a further 20 patent applications pending. He was a member of the security team of IBM Research (Zurich) and the Director of IBM's Privacy Research Institute. Currently Thomas investigates cloud security assurance (EU H2020 PrismaCloud, GA no 644962. Thomas has trained as a psychologist and is investigating human dimensions of cyber security following on from his Co-Investigator role in the EPSRC Research Institute in the Science of Cyber Security (RISCS). Thomas was the Principal Investigator of a recent RISCS II project on evidence-based methods in cyber security as well as the Principal Investigator of the Newcastle Human Dimensions in Cyber Security Lab.

Suggested keywords

cyber security, user study, systematic literature review, SLR, evidence-based methods

A Systematic Evaluation of Evidence-Based Methods in Cyber Security User Studies

Kovila Coopamootoo
Newcastle University
Newcastle upon Tyne, UK

Thomas Groß
Newcastle University
Newcastle upon Tyne, UK

***Abstract*—Background. In the recent years, there has been a movement to strengthen evidence-based methods in cyber security under the flag of "science of security." It is therefore an opportune time to take stock of the state-of-play of the field.**
Aim. We evaluated the state-of-play of evidence-based methods in cyber security user studies.
Method. We conducted a systematic literature review study [1] of cyber security user studies from relevant venues in the years 2006–2016. We established a qualitative coding of the included sample papers with an *a priori* codebook of 9 indicators of reporting completeness [2]. We further extracted effect sizes for papers with parametric tests on differences between means for a quantitative analysis of effect size distribution and post-hoc power.
Results. We observed that only 21% of studies replicated existing methods while 78% provided the documentation to enable future replication. With respect to internal validity, we found that only 24% provided operationalization of research questions and hypotheses. We observed that reporting did largely not adhere to APA guidelines as relevant reporting standard [3]: only 6% provided comprehensive reporting of results that would support meta-analysis. We, further, noticed a considerable reliance on p-value significance, where only 1% of the studies provided effect size estimates [4]. Of the tests selected for quantitative analysis, 80% reported a trivial to small effect, while only 28% had post-hoc power ($1 - \beta \geq 80\%$). Only 16% were still statistically significant after Bonferroni correction for the multiple-comparisons made.
Conclusions. This study offers a first evidence-based reflection on the state-of-play in the field and indicates areas that could help advancing the field's research methodology.

***Index Terms*—cyber security, user study, systematic literature review, SLR, evidence-based methods**

I. Introduction

The Encyclopaedia Britannica defines *science* as a "system of knowledge that is concerned with the physical world and its phenomena and that entails unbiased observations and systematic experimentation." The scientific method includes principles such as falsification of hypotheses or reproducibility as well as statistical tools to decide between hypotheses.

A recent movement sought to strengthen evidence-based methods in cyber security under the flag of "science of security." We believe it is an opportune time to take stock of the state-of-play observed in the field.

As an inter-disciplinary research domain, cyber security benefits from inputs from a number of sciences. There have been a number of proposals on how to improve the quality of evidence-based research in security and privacy, from making experiments dependable [5], over guidance of how to conduct experiments in security and privacy [6], [7], avoiding pitfalls when writing about security and privacy security experiments [8], to introductions to evidence-based methodology [7].

While those guides offer evidence for reflection in the field, we perceive that they are either founded on hallmarks of scientific research or anecdotal evidence of problems observed by program committees. We find that a systematic evaluation of "how the field is actually doing" has never been attempted at scale.

The aim of this study is to fill this research gap with a systematic literature review of cyber security user studies in the same timeframe that most of said guides were published (2006–2016). We pursue a qualitative coding of reporting completeness based on an *a priori* codebook of 9 completeness indicators [2]. Our research questions and the corresponding codebook cover (i) up- and downstream replication, (ii) internal validity of the studies, incl. explicit formulation of research questions and hypotheses, (iii) adherence to APA reporting standards, (iv) reporting of effect sizes, (v) overall soundness of the results. The codes operate to a large extend syntactically on whether certain pieces of information are present or absent.

With respect to the quantitative soundness of the results we extract a sub-sample with parametric tests of differences between means to evaluate effect sizes and post-hoc power.

a) Contributions: In the remaining of this paper, we present a first systematic literature review of the state of field for human factors of cyber security that does not only account for quality criteria such as study validity or indicators of reproducibility, but also give a quantitative measurement of the impact of the corresponding studies.

b) Outline: In the rest of the paper, we provide the research aim and a detailed methodology for both the qualitative part involving a review process for Completeness Indicators and the quantitative meta-analysis. We then provide our findings followed by a discussion and conclusion. In the Appendix, we provide the list of 146 papers that were part of the systematic literature review. We opt to not provide a related work section since the rest of the paper grounds research decisions in a comprehensive foundation.

II. Aim

We aim to gather and summarize evidence concerning the state-of-play of user studies in cyber security through a systematic literature review.

a) Research Questions.: We define research questions seeking to evaluate whether the study under evaluation addressed the hallmarks for scientific research. In particular, we asked

- RQ1 Did the experiment repeat or reproduce existing studies/methods? Was the experiment sufficiently reported to enable reproducibility?
- RQ2 To what extent were the described studies internally valid?
- RQ3 How many of the eligible papers reported results from experiments correctly according to APA guidance?
- RQ4 To what extent were effect sizes and power estimates provided? How many of the studies had appropriate power?
- RQ5 How many of the results reported agree with an independent recalculation of test statistics and effect sizes?

III. Method

A systematic review aims to synthesize existing research in a manner that is fair. At the same time, a well defined methodology makes it less likely that the results synthesis of the literature are biased. When consistent results are observed across studies, the systematic review provides an indication whether a phenomenon is robust, and a meta-analysis enables detection of real effects that individual studies can miss out on. We designed a survey following systematic literature review guidelines proposed by Kitchenham [1].

The systematic review consists of three main parts: a search process that identifies primary studies addressing the research questions, a data extraction process that extracts the data items needed to answer the questions and a data analysis process that synthesizes the data.

A. Procedure

A systematic literature review follows a predefined search strategy, a review protocol, that specifies the methodology to undertake to conduct the systematic review. (a) First we start with research questions that the review is intended to answer. (b) We then define a strategy that aims to detect relevant literature. (c) We set study selection criteria which determine which study are included in or excluded from the review. (d) We then specify the information to be obtained from each study, including a set of quality assessment checks (Completeness Indicators) to assess individual studies.

(e) We decide how the information required from each study is obtained in a data extraction strategy. (f) And lastly we synthesize the extracted data.

B. Search Query

Since one aim of the systematic review is to find as many primary studies relating to the research question as possible using an unbiased search strategy, the rigor and completeness of the search query is vital. We setup a search strategy that include papers from 10 years (2006–2016) with a source list from:

- journals, such as IEEE Transactions on Dependable & Secure Computing (TDSC), ACM Transactions on Information and System Security (TISSEC),
- flagship security conferences, such as IEEE S&P, ACM CCS, ESORICS, and PETS or
- specialized venues, such as LASER, SOUPS, USEC and WEIS.

We define our search query on Google Scholar. Each query extracts articles mentioning "*user study*" and at least one of the words "*experiment*", "*evidence*" or "*evidence based*". We run the query for each of the 10 publication venues. In the advanced search option of Google Scholar, we set each of the following fields:

- with all words = *user study*
- at least one of the words = *experiment evidence "evidence based"*
- where my words occur = *anywhere in the article*
- return articles published in = [publication venue]
- return articles dated between = *2006 — 2016*

Consequently, we extracted 1157 articles spread across the 10 venues as shown in Figure 1 and Table I.

C. Inclusion and Exclusion Criteria

From the search query results, we focus the systematic review on human factors studies including a human sample. Of the pool of 1157 articles, papers fulfilling the following *Inclusion Criteria* were included:

- Studies including a user study with human participants.

Table I: # Articles extracted by publication venue

Venue	N
Learning from Authoritative Security Experiments Results (LASER)	07
Workshop on the Economic of Information Security (WEIS)	09
Usable Security (USEC)	12
ACM Transactions on Information and System Security (TISSEC)	76
Symposium on Usable Privacy & Security (SOUPS) in full	168
Symposium on Usable Privacy & Security (SOUPS) acronym	91
Privacy Enhancing Technologies Symposium (PETS)	99
IEEE Symposium on Security & Privacy (IEEE S&P)	121
IEEE Transactions on Dependable & Secure Computing (TDSC)	161
USENIX Security	197
ACM Conference on Computer and Communications Security (CCS)	216

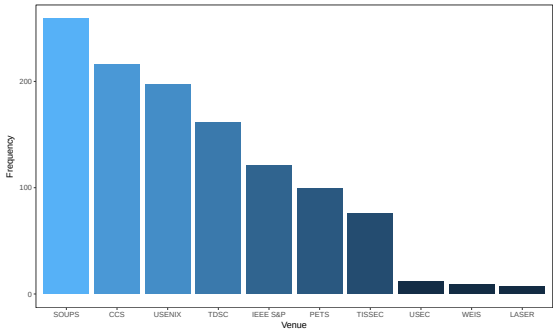


Figure 1: Spread of articles across publication venues.

- Studies concerned with evidence-based methods or eligible for hypothesis testing and statistical inference.
- Studies that lend themselves to quantitative evaluation, quoting statements of statistical significance, p-values or effect sizes.
- Studies with true experiments, quasi-experiments or observational analysis.

Of the papers included, the ones matching these *Exclusion Criteria* were excluded:

- Papers that were not subject to research peer-review, key note statements, posters and workshop proposals.
- Position papers or informal arguments.
- Papers not including a study with human participants,
- Theoretical papers.
- Studies with qualitative methodology.

After filtering the articles via the inclusion and exclusion, we end up with a total of 146 articles, that we provide in the Appendix. Of these, we later

find that 112 of the articles were eligible for full Completeness Indicator evaluation while 19 were eligible for meta-analysis. Figure 2 summarizes the process. The 34 articles not eligible for Completeness Indicator evaluation included those without a user study involving human participants, without p-value statistics or with sequential equation modeling or support vector machine for machine learning rather than Null Hypothesis Significance Testing (NHST) or effect size/confidence interval estimation.

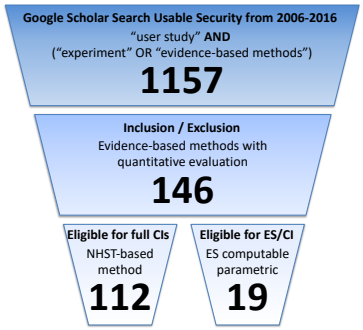


Figure 2: Search query & refinements on sample.

D. Completeness Indicators.

A systematic review is conducted, in which a set of nine Completeness Indicators (CIs) is coded. We focus on Completeness Indicators derived from the 5 research questions defined in Section II-0a. In figure 3 we show how the 9 CIs pertain to the 5 research questions.

We opted for research questions and CIs founded on the hallmarks for empirical research and statistical inference, that are easily observable in the article reviewed. We left out research criteria such as external validity, noting the trade-off between internal and external validity, that is that not all experiments can be internally and externally valid at the same time. Furthermore, it depends whether the purpose of a particular study seeks high internal or high external validity, where typically one kind of validity is sacrificed for another [5].

As data extraction strategy, we code the papers across the completeness indicators (CIs) in NVivo,

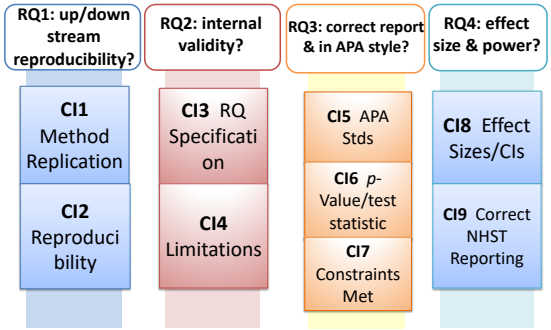


Figure 3: The CIs derived from the research questions.

extracting properties defined in the sub-criteria for each CI described below.

In short, the CIs are:

- CI1 Was the study replicating existing studies or methods?
- CI2 Was there correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method, contributing towards reproducibility?
- CI3 Was there an explicit and operational specification of the RQs, null and alternative hypotheses, IVs, DVs, subject assignment method and manipulation checks?
- CI4 Was there a discussion on the limitations, possible confounders, biases and assumptions made?
- CI5 Was the result reported in the APA style?
- CI6 Did the result statement include test statistic and p-value?
- CI7 Were significance level α and test statistics properties and assumptions appropriately stated (e.g., "two-tailed")?
- CI8 Were the appropriate the effect sizes and confidence intervals (CI) reported?
- CI9 Was the significance and hypothesis testing decision interpreted correctly and put in context of effect size and sample size/power?

E. CI1: Was the study replicating existing studies or methods?

Research claims gain credence when the supporting evidence can be replicated [9]. Replicability is an important practice for *Open Science* who notes the alarming discovery that a number of widely known and accepted research findings cannot be replicated [4]. Subsequently Cumming & Calin-Jageman [4] point out that rarely, if ever can a single finding give definitive answer to a research question, making replication important for confidence in the finding.

In the recent years, the security community, including SOUPS, has been encouraging replication of existing studies. In security literature, Maxion [5] postulates that repeatability, reproducibility and validity are the main criteria differentiating a well designed experiment from those that are not. In other sciences such as psychology, a replication crisis has already been observed. A large scale replication endeavor by the Open Science Foundation [9] of $N = 100$ studies across 3 psychology journals found that only 47% of the original effect sizes were in the 95% confidence interval of the replication effect size.

While Coopamootoo & Groß [7] distinguish between repeatability and reproducibility as two conceptual frames for research replication, we extend the conceptualization to (a) *upstream replication* where previous studies or validated methods are replicated, versus (b) *downstream replication* where the study is sufficiently reported and is thereby reproducible by other researchers or research groups. In the SLR, we code for upstream replication, that is whether the reviewed study built on solid and validated foundations, replicating existing studies de-facto or replicate previously validated manipulation and measurement apparatus, or developed adaptations. Table II shows our coding for partial or complete fulfillment of this CI and for failure, where the impact of CI1 is the assurance that the study was designed from solid foundations.

Table II: Criteria for CI1. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Fail
Replicated existing methods as is	Yes		No
Adapted existing methods		Yes	No

F. CI2: Was there correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method, contributing towards reproducibility?

We evaluate whether the article gave sufficient reports of the study to enable downstream replication, that is whether (another researcher) could possibly reproduce the study given the detailed study report. Table III shows our coding for partial or complete fulfillment of this CI and for failure. The impact of CI2 is that it enables downstream replication and reuse of methods.

Table III: Criteria for CI2. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Partial	Fail
Measurement and manipulation apparatus	Yes	No		No
Detailed Procedure	Yes			No
Sample Size	Yes			
Demographics	Yes		No	
Sampling and Recruitment	Yes			

G. CI3: Was there an explicit and operational specification of the RQs, null and alternative hypotheses, IVs, DVs, subject assignment method and manipulation checks?

While validity refers to whether the experiment is actually measuring what was intended [10], internal validity refers to the truth that can be ascribed to cause-effect relationships between independent variables (IV) and dependent variables (DV) [11], where the IV is a variable that is induced/manipulated and the DV is the variable that is observed/measured [12]. In the SLR, we coded research questions and hypotheses that provide the foundations for null hypothesis significance testing (NHST) [13], such as as guided in [7]. Operationalization enables systematic and explicit clarification of the predictors, IVs, and hence the

cause and manipulation, while the target variable or DVs clarifies the effect, hence the measurements. Subject assignment points to whether and how participants were randomly assigned and balanced across experimental conditions hence avoiding a bias and other possible explanations for between-subject designs. For within-subject studies, random assignment to manipulation sequences counters order effects. Manipulation check refers to verification that the manipulation has actually taken effect, hence lowering possible doubts that the observed effect did not emanate from the induced manipulation. Table IV shows our coding for partial or complete fulfillment of CI3 and for failure. The impact of CI3 is ensuring internal validity and a solid statement of intention for Null Hypothesis Significance Testing (NHST).

Table IV: Criteria for CI3. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Partial	Partial	Fail
Research Question	Yes				No
Hypotheses	Yes			No	No
IVs and DVs	Yes			No	No
Subject Assignment	Yes	No			
Manipulation Check	Yes		No		

H. CI4: Was there a discussion on the limitations, possible confounders, biases and assumptions made?

A discussion of the limits and boundaries of the study, identification of possible confounding variables whose presence affect the relationship under study, and possible assumptions made in setup, are all valuable inputs that strengthen the validity of the experiment. Table V shows our coding for partial or complete fulfillment of CI4 and for failure. The impact of CI4 is transparency of validity and other possible explanations for the stated causal relations.

Table V: Criteria for CI4. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Partial	Partial	Fail
Research Limitations	Yes	No			No
Confounders	Yes		No		No
Biases (sampling)	Yes			No	No

I. CI5: Was the result reported in the APA style?

Reporting standards provide a degree of comprehensiveness in the information that is reported for empirical investigations. Uniform reporting standards make it easier to generalize within and across fields, to understand implications of individual studies and to allow for techniques of meta-analysis. Comprehensive reporting also supports decision makers in policy and practice towards understanding how the research was conducted [3]. We elect to ask for the reporting recommendations of the American Psychology Association (APA) [3] as quality standard. The APA provides specific guidelines for reporting statistical results [3]. Table VI shows our coding for partial or complete fulfillment of CI5 and for failure. The impact of fulfilling CI5 is a standardized form of reporting as a driver for research quality, reuse and reproducibility.

Table VI: Criteria for CI5. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Fail
APA guidelines for all results	Yes		No
APA guidelines for some results		Yes	No

J. CI6: Did the result statement include test statistic and p-value?

This CI supports reproducibility of the analysis and foundations for research evidence and quality.

Table VII shows our coding for partial or complete fulfillment of CI6 and for failure. The impact of CI6 is foundation for post-hoc analysis and multiple-comparisons corrections.

Table VII: Criteria for CI6. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Fail
Actual $p-value$ reported	Yes	No	No
Test-statistics reported	Yes	Yes	No
Mean & standard Dev. reported	Yes		

K. CI7: Were significance level α and test statistics properties and assumptions appropriately stated ?

To ascertain whether the statistical analyses were correctly employed on the data, statistical assump-

tions need to be made explicit in reporting. For example, the assumptions for parametric tests, in general, are normally distributed data, homogeneity of variance, interval data and independence [14]. Example for test properties is "one-tailed" or "two-tailed". Table VIII shows our coding for partial or complete fulfillment of CI7 and for failure. The impact of CI7 is proof of appropriateness and correct deployment of the statistical methods used.

Table VIII: Criteria for CI7. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Partial	Partial	Fail
Significance level	Yes			No	No
Test assumptions	Yes		No		No
Test Properties	Yes	No			No

L. CI8: Were the appropriate the effect sizes and confidence intervals (CI) reported?

An effect that is statistically significant is not necessarily scientifically significant or important, where the importance of an effect is linked to the magnitude of the effect [15]. In addition, the APA makes reporting of confidence intervals a minimum standard.

Kirk [16] and Cumming [17] debated that the current research practice of exclusive focusing on a dichotomous reject-nonreject decision strategy of null hypothesis testing that can impeded scientific progress. Rather, they posit, the focus should be on the magnitude of effects, that is the practical significance of effects and the steady accumulation of knowledge. They advise to switch from the much disputed NHST to effect sizes, estimation and cumulation of evidence. In the estimation approach to inferential statistics, the effect size (ES) provides a *point estimate* of support in the population while the confidence interval (CI) provides the interval estimate, whose length indicates the precision of estimation. The 95% CI provides confidence that the true value of support in the population lies in the interval estimate. A short CI points to a small margin of error and a relatively precise estimate that the point estimate is likely close to the population value whereas a long CI means a large margin of error and low precision.

This approach is also supported by the APA guidelines [3], that states that "estimates of appropriate effect sizes and confidence intervals are the minimum expectations." That implies to make effects and coefficients of regressions available. CI8 includes that the effect sizes are reported in a easily interpretably form. Table IX shows our coding for partial or complete fulfillment of CI8 and for failure. The impact of CI8 is parameter estimation as robust report of effect magnitude and foundation for meta-analysis and cummulation of knowledge.

Table IX: Criteria for CI8. Yes = Present, No = Absent.

Sub-criteria	Success	Partial	Partial	Fail
Effect sizes for all results	Yes			
Effect sizes for some results			Yes	No
Confidence intervals	Yes	Yes		No

M. CI9: Was the significance and hypothesis testing decision interpreted correctly and put in context of effect size and sample size/power?

Nickerson [13] offers a comprehensive overview of the controversies around *Null Hypothesis Significance Testing (NHST)*, while Maxwell and Delaney [18, p.48] and Goodman [19] point to p-Value misconceptions and Ioannidis [20] argues "why most published research findings are false." The misconceptions around NHST include the beliefs that [13]

- p is the probability that the hypothesis is true and $1 - p$ the probability that the alternative hypothesis is true,
- a small p is evidence that the results are replicable,
- a small value of p means a treatment effect of large magnitude,
- statistical significance means theoretical or practical significance,
- alpha is the probability that a Type I error will be made,
- beta is taken to mean the probability that the null hypothesis is false,
- failing to reject the null hypothesis is equivalent to demonstrating it to be true,

- failure to reject the null hypothesis is evidence of a failed experiment

In addition, Ioannidis presents some corollaries supporting the argument of why most published research are false [20]

- the smaller the sample size of the study,
- the smaller the effect size,
- the greater the tested relationships and the lesser the selection of tested relationship,
- the greater the flexibility in design and analytical modes,
- the greater the financial interests,
- the hotter the scientific field,

the less likely the research findings are true.

CI9 asks for correctness in how statements on statistical significance are expressed and what conclusions are drawn from the statement. This includes Cohen's creed [15] that significance needs to to be considered vis-à-vis of sample size and power of the experiment.

N. Quantitative Analysis

The quantitative analysis aims primarily at evaluating effect sizes and 95% confidence intervals independently. We focus on effect sizes from parametric tests, that is, mostly differences between mean in normally distributed data. We choose Hedges' g as effect size metric for standardized mean differences as an unbiased effect size favored in meta analysis. We note that only a minority of papers report effect sizes explicitly and only 46% reported sufficient data to infer Hedges' g.

We depict the analysis workflow in Figure 4. The first stage involves coding the papers and their properties in NVivo. This coding involves the specification of samples, effect sizes and relations tested as well as properties, such as the use of an MTurk sample and the correction for multiple comparisons. Papers that are based on non-parametric tests or do not give sufficient information to obtain Hedges' g are discarded at this stage.

The data used to compute effect sizes and confidence intervals thereon is then transferred to R for the quantitative analysis, using packages meant for meta-analysis (metafor) and parameter estimation (MBESS).

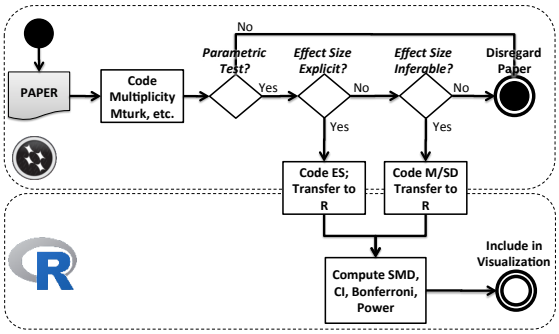


Figure 4: Workflow of the quantitative analysis towards effect sizes, power, and confidence.

We then compute

- standardized mean differences (Hedges' g),
- the post-hoc power at a significance level $\alpha = .05$,
- the 95% confidence interval on the effect size,
- the corresponding margin of error, as well as
- the margin of error with a per-study Bonferroni correction.

To compute this Bonferroni correction, we count the number of parametric comparisons made within a study and adjust the study's significance level α by the number of comparisons.

IV. Results

A. General

a) Security Research Theme: We found that 32 of the 112 articles addressed privacy research whereas 26 were on password authentication. The rest of the articles were spread across a variety of security themes with smartphone security and warning and dialogs taking the next chunks, each with a count of 7.

b) Publication Venue: From the 112 articles, we found that 76 were from SOUPS (making 68%), and the rest spread across the different venues as shown in Table X.

B. Completeness Indicators

For the Completeness Indicator evaluation, we analyzed the 112 articles, making a total of 134

Table X: # Articles reviewed by publication venue

Venue	N
Learning from Authoritative Security Experiments Results (LASER)	02
Workshop on the Economic of Information Security (WEIS)	01
Usable Security (USEC)	04
ACM Transactions on Information and System Security (TISSEC)	02
Symposium on Usable Privacy & Security	76
Privacy Enhancing Technologies Symposium (PETS)	08
IEEE Symposium on Security & Privacy (IEEE S&P)	01
IEEE Transactions on Dependable & Secure Computing (TDSC)	03
USENIX Security	08
ACM Conference on Computer and Communications Security (CCS)	07

studies. Figure 5 shows an overall view of the results.

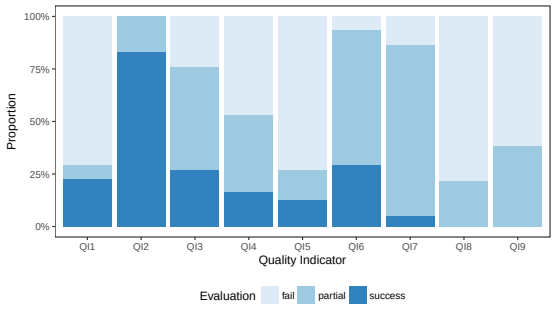


Figure 5: Evaluation results across CIs.

1) QI1 - Upstream Replication: We found that 74% of the studies did not replicate an existing measurement method nor a whole study whereas 21% did and 5% adapted an existing method, as depicted in Figure 6.

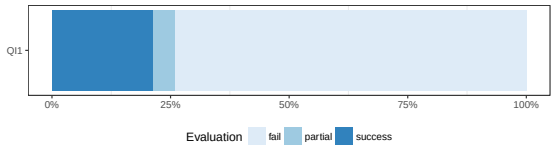


Figure 6: Evaluation results for CI1.

2) QI2 - Research Reproducibility: To enable a research study to be reproduced in the future, we evaluated for correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method. We found that 20% of the studies was not documented enough to enable reproducibility, 78% did while 2% was not complete, as in Figure 7.

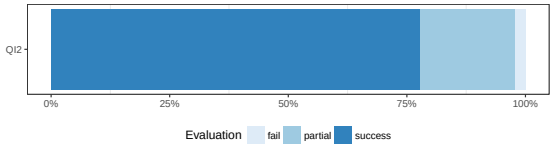


Figure 7: Evaluation results for CI2.

3) QI3 - Internal Validity - Operationalization of Hypotheses: We asked whether the studies described are internally valid by examining whether they specified Research Questions, null and alternative hypotheses, Independent Variables, Dependent Variables, subject assignment method and manipulation checks. Based on the information provided in the articles, we found that for 26% of the articles, there was not enough information to support the validity of the study while 24% were clearly valid, and 50% only partially, as in Figure 8.

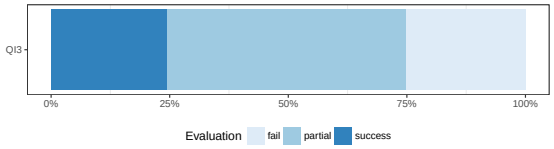


Figure 8: Evaluation results for CI3.

4) QI4 - Limitations: Another aspect of evaluating validity of the studies is through an assessment of the biases, confounders and limitations of the study. We found that 48% did not provide a discussion of the limitations of the study, while 18% did and 34% not covering all the components of this CI, as shown in Figure 9.

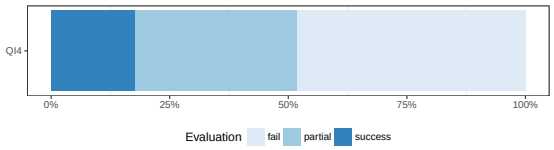


Figure 9: Evaluation results for CI4.

5) QI5 - Standard Reporting: We evaluated whether the studies reported their results according to the APA guidelines [3]. we found that only 13% of the studies did so completely, while 16% provided standard reports for some results only and 71% did not adhere to standard reporting guidelines. See Figure 10 for a depiction.

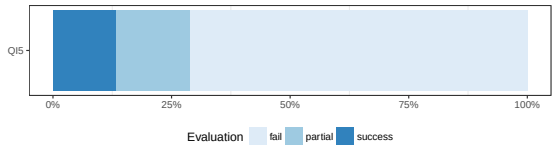


Figure 10: Evaluation results for CI5.

6) QI6 - Test Statistic & p-value: We found that 31% of the studies reported the actual p-value, the test-statistics and the means and standard deviations. 62% provided the test statistic without the actual p-value and 8% failed to report either of them, as shown in Figure 11.

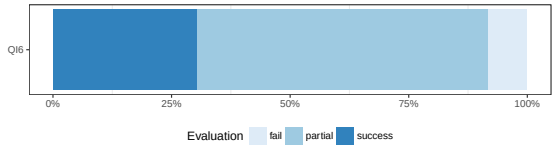


Figure 11: Evaluation results for CI6.

7) QI7 - Alpha level, test assumptions and properties: We found that 75% of the studies failed to provide the complete set of sub-criteria for this CI, that is they either missed out on significance level, test assumptions or test properties. Only 6% of the studies provided the complete set, while 19% failed completely. See Figure 12 for a depiction.

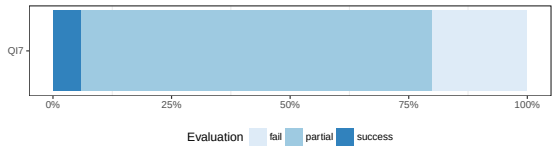


Figure 12: Evaluation results for CI7.

8) QI8 - Effect Size & Confidence Intervals: We found that only 1% of the studies provided effect sizes for all results and their confidence intervals. 20% reported either effect sizes for some results or confidence intervals only, while 79% did not provide effect sizes nor confidence intervals, as shown in Figure 13.

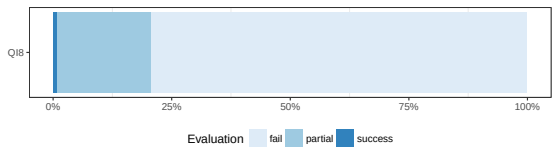


Figure 13: Evaluation results for CI8.

9) QI9 - NHST interpretation: We found that only 1% of the studies provided a correct p-value interpretation, a-priori sample specification, Type-I error correction, specification of null and alternative hypotheses and population specification. 35% provided only p-value interpretation and either of Type I error correction or the alternative hypotheses. 64% failed this CI by providing only the p-value. See Figure 14 for a depiction.

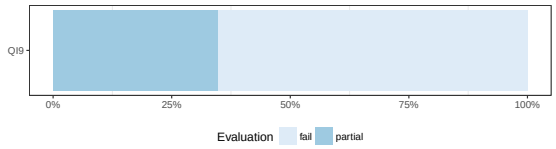


Figure 14: Evaluation results for CI9.

C. Quantitative Analysis

In the quantitative analysis we focused on papers, which used parametric statistics and for which we could derive the standard mean difference in Hedges' g from the data presented in the paper (e.g., means and standard differences). In this part we considered 19 papers, which fulfilled both constraints.

The given $n = 19$ papers made comparisons on 277 relations in total, 148 comparisons with a non-trivial effect size in Hedges' $g > 0.2$.

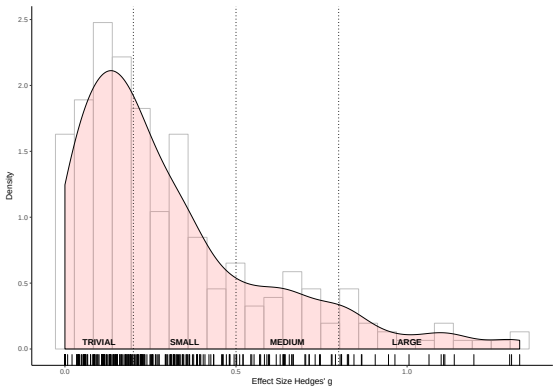


Figure 15: Density of observed effect sizes (based on 277 mean differences from $n = 19$ papers).

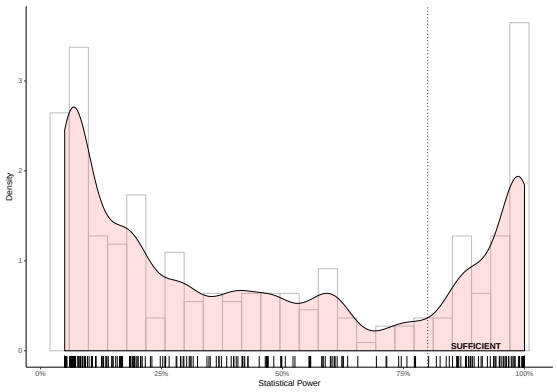


Figure 16: Density of observed power (based on 277 mean differences from $n = 19$ studies).

1) Effect Sizes: We computed the point estimates for all comparisons made in the selected sub-sample. As adequate point estimate for meta-analysis on studies with differences between means, we selected Hedges' g. This standard mean difference bears the advantage that it is an unbiased effect size. Figure 15 shows the distribution of effect sizes in a density plot overlaid on a histogram. As a rule-of-thumb guidance, we denote areas of the distribution that are considered "trivial," "small," "medium," or "large" by Cohen's classification.

47% of the effects in observed relations were trivial, 33% were small, 13% were medium, 8% were large. Of all relations considered in the $n = 19$ studies, 60% were not statistically significant (without correction for multiple comparisons made). Of the $112(40\%)$ statistically significant effects in the sample, 59% had a small magnitude, 21% had a medium magnitude, 20% had a large magnitude.

2) Statistical Power: Given the effect sizes observed and the sample sizes of the respective studies, we gain an estimate on the post-hoc statistical power of the studies. We visualize the power distribution over all comparisons made of the respective studies in Figure 16. The usual recommendation for sufficient statistical power is $1 - \beta = 80\%$.

Of all the relations investigated by the selected sample, 38% had negligible power ($1 - \beta < 20\%$), 21% had a small post-hoc power ($20\% \leq 1 - \beta < 50\%$), 12% had medium power ($1-\beta < 80\%$), 29% had sufficient power ($1 - \beta \geq 80\%$).

3) Confidence and Margins of Error: We considered two kinds of 95% confidence intervals on the effect sizes (Hedges' g): the primary confidence interval is on the effect size without corrections. The second confidence interval considers the the per-study number of comparisons and employs a Bonferroni correction on the confidence interval, even if the authors of the respective study did not correct for multiple comparisons made.

From the confidence intervals, derive the margin of error (half the length of the confidence interval). We display the distribution of margins of error in percent of the corresponding Hedges' g in Figure 17, where we restrict the visualization to effects of Hedges' $g > 0.2$.

Subfigure 17a shows the margins of error without correction of multiple comparisons. This view is the most optimistic view on the margins of error for the comparisons made. In this case, most common margin of error is 60%. If we were to consider an effect of 1 *SD* standardized mean difference, Hedges' $g = 1 \pm 60\%$, the 95% CI would be $[0.4, 1.6]$.

Subfigure 17b adjusts the confidence intervals and corresponding margins of error with a per-study Bonferroni correction. Under per-study Bonferroni correction, only 16% of the comparisons made are still statistically significant (compared to 40%

without correction).

V. Discussion

A. There Is a replication crisis.

From CI1 and CI2, we observe that there is a replication crisis in the research area of human factors of cyber-security. While a large portion (78%) of the studies reviewed provided enough details to enable future reproduction, only 21% replicated existing methods, that is measurements and manipulations that have previously been tested by others.

There was only 1 reported replication of an existing study with small enhancements. In consequence, similar to replication crises previously observed in other research fields such as psychology [20], cyber-security research is currently facing such a problem.

Without replication, it is rare, if ever, possible to determine if the findings of a single study is definitive. Close replications often provide additional evidence which with meta-analyses contribute more precise estimates [17]. Studies that keep some original features and vary others can also offer a converging perspective. In addition, scientific claims gain credence when their supporting evidence can be replicated [9]. Therefore, to benefit from research evidence that have the potential to influence policy and practice, the cyber-security research community ought to encourage and perhaps even provide incentives for research replications.

B. Internal validity need to be called in question.

Validity refers to the best possible approximation to the truth and falsity of propositions [21]. Hence validity ensures an argument is logically correct, sound and flawlessly reasoned. Internal validity refers to the truth that can be assigned to the conclusion that a cause-effect relationship between an *IV* and a *DV* has been established [11]. From CI3, we observe that only 24% of the studies specified hypotheses, operationalized into variables hence enabling evaluation of the internal validity of the study. Therefore, while experts in cyber-security experimentation postulate that we need to ensure that measurements are dependable and error-free [5], the observations of the current SLR point to a problem in the field. For example, a critical form of error in experiment designs is the confound – where the value of a variable is confounded or influenced by the value of another. We observe from CI4, that only 18% of the studies provided a discussion of limitations, with a smaller percentage addressing confounders.

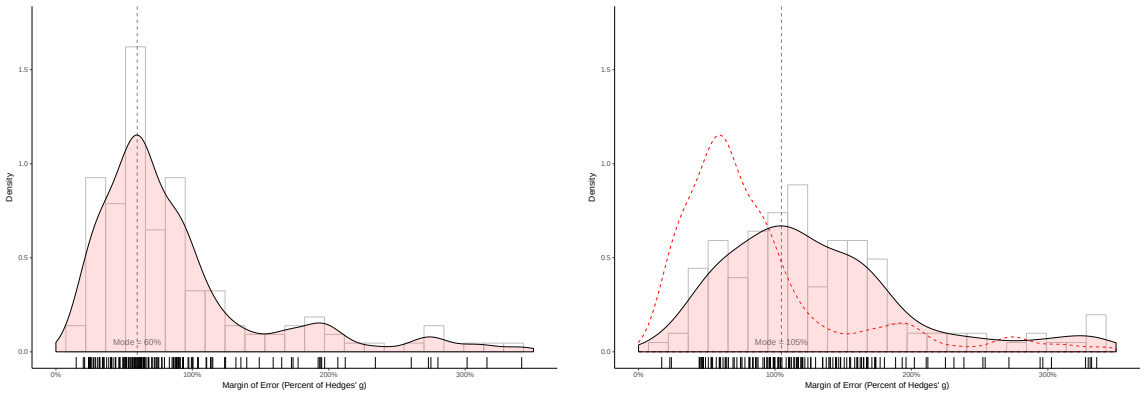
The problems emerging from such low assurance is whether the researchers can rely on the results, where the possibility to 'stand on the shoulders of giants' and contribute to the progress of scientific knowledge in cyber-security is impeded.

C. The field would benefit from standardized statistical reporting.

Standardized reporting provides a degree of comprehensiveness in the reported information, makes it easier to generalize within and across fields, to understand the implications of individual studies and to allow for meta-analysis that in turn supports scientific credibility [3]. We observed that only 13% of studies adhered to the APA guidelines CI5, 31% reported on the actual *p-value* (CI6) and 6% provided statistical test assumptions and properties (CI7). As additional evidence, we found that only 19 of the 146 articles were eligible for meta-analysis, that is provided enough information to determine effect sizes and statistical power.

D. Reliance on p-value.

Effect size estimate provide an indication of the magnitude of the observed effect [15], [22], hence helping to distinguish between effects that are trivial or negligible from effects that are likely making a difference in real-world applications. Confidence intervals provide an interval estimate that the true value of the population effect size lies in the estimated interval [4]. From CI8, we observe that only 1% of the studies provided effect sizes and confidence intervals for all their results. Hence we observe a reliance on significant *p-values*, an approach already much disputed in literature [20], [19], [4]. From CI9, we also observe that only 1% of the studies provided enough information to ascertain correct evaluation of hypotheses.



(a) Uncorrected.

(b) With per-study Bonferroni correction.

Figure 17: Margin of errors on effect sizes without and with per-study Bonferroni correction (based on 148 mean differences with effect sizes Hedges $g > 0.2$). The distributions are plotted on the same scale, where the dashed red line denotes the uncorrected distribution.

E. Small observed effect sizes.

We perceive that the majority of effects reported as statistically significant in quantitative sub-sample of $n = 19$ papers had small effects (59%). Even though these effects are statistically significant, they might not be practically relevant or scientifically significant.

This observation is made vis-à-vis of a low rate of explicitly reporting effect sizes at all (8%) or having sufficient data in the paper to derive standardized effect sizes for subsequent meta-analysis.

We find that evidence-based papers on user studies in security and privacy are largely focused on reporting statistically significant results, ignoring parameter and interval estimation. This is troublesome news, especially, considering that many of the effects reported as significant are actually small.

From these observations, we strongly recommend to authors and program committees alike to adopt parameter and interval estimation [17], that is, to report standardized effect sizes and confidence intervals on them. We recommend to use unbiased effect sizes that are not easily derived from other data, such as t-values.

F. Observed power.

From the quantitative sub-sample ($n = 19$), we observe that less than one third of the comparisons were made at adequate power by rule-of-thumb considerations ($1 - \beta \geq 80\%$). By and large, the majority of studies investigated seemed underpowered, with the notable exception of studies with larger samples drawn from Amazon Mechanical Turk (AMT).

While the low power means in first instance that the change of Type II errors is greater (rejecting the alternative hypothesis, when it is true in reality), the low power has further consequences. Specifically, the Positive Predictive Value (PPV) [20] will be lower, which means that the reported results are more likely false positives. Conversely, the studies are less likely to achieve noteworthiness ($PPV \geq 80\%$) [23]. Hence, in addition to reducing the likelihood of finding our about real effects, the studies' outcomes are also less trustworthy.

We recommend for evidence-based user studies in security to assure adequate power, by either conducting an *a priori* power analysis [22] or by going a step further in endorsing Accuracy in Parameter Estimation (AIPE) [24]. With such methods, the experimenters need to obtain a sound estimate on the effect sizes during the design phase (either from

the literature or pre-tests) and compute required adequate sample sizes from them.

G. Margins of error.

Even without correction for multiple comparisons, the most common margin of error on effects observed was 60%. Hence, such studies could only offer an interval estimate (95% Confidence Interval) on their respective effect size in Hedges' $g \pm 60\%$.

Hence, we observe that, generally, the studies in the quantitative sub-sample ($n = 19$) were not able to yield a tight confidence interval on the effects observed. In turn, this means we have little certainty on how large the effect in the population might be.

Under the consideration of per-study corrections for multiple comparisons, the margins of errors are considerably greater, calling the results further into question.

First, it is important to raise awareness in the community that statistically significant effects do not necessarily also mean reliable effects. The Accuracy in Parameter Estimation (AIPE) [24] mentioned previously offers assurances that a study will be adequately powered to gain tight confidence intervals at sufficient confidence.

H. Limitations

a) Completeness Indicators and Qualitative Coding.: The completeness indicators defined in the codebook [2] have a limited scope. While they aim at ascertaining up- and downstream replication, reporting supporting internal validity and adhering to standards, as well as aspects of quantitative reporting, they are take largely syntactic snapshots of the studies in question. We stress that such indicators only capture face validity and do not penetrate the inner argument of the studies deeply.

Ideally such indicators would be complemented with well-evidenced codebooks for random-controlled trials [25] (e.g. the well-known Jadad scale.) or auxiliary reporting spearheaded in open science (e.g., pre-registrations, published materials and detailed study protocols, account for all tests computed).

b) Generalizability.: This study focuses on user studies in cyber security from 2006–2016. We found that the search on Google Scholar was somewhat hit-and-miss: A number of papers found were not actually user studies. We need to assume that, similarly, the search missed studies that would have been considered valid for inclusion.

In the included sample itself we also found a cross-section of different types of papers, most notably studies who focused on human factors/user studies as main line of inquiry and studies who had a small user study tagged on, outside of the primary line of inquiry of the study.

This may make for a faithful representation of the situation in the field, at the same time, we believe the generalizability of the SLR to be limited due to the properties of the sample.

c) Quantitative Results.: We note that the quantitative analysis considered observed, that is, *post-hoc*, effect sizes, their confidence intervals and power. Post-hoc power is redundant with the reporting on p-values itself.

We are aware that the post-hoc analysis could fall for flukes and overestimate effect sizes and power alike. The reason for that is that low-power studies are also more likely to report false positive results and over-estimated effect sizes, which in turn foils the post-hoc power analysis to over-estimate the power achieved. Furthermore, the post-hoc power analysis does not account for power lost due to decisions made by the authors during their study.

In spite of the uncertainty introduced by the post-hoc estimation, we believe the analysis still offers a glimpse at the power situation found in the studies.

VI. Conclusion

We provide a first systematic review of cyber-security user studies. It offers a wealth of insights in the state-of-play of the field as well as pointers on how to improve the situation.

We offered evidence that few of the studies build on validated tools or replicate existing methods. We consider that a replication crisis as such studies are unlikely to be replicable. We saw challenges in accepting face internal validity and concluded that the field would strongly benefit from reporting

standards. Overall the effects observed were small, often too small to have a practical effect, stressing the importance of ascertaining the magnitude of the effects, not just significance.

ACKNOWLEDGMENTS

This research was first made public in the community meeting of the UK Research Institute in the Science of Cyber Security (RISCS), in February 2017. A first version of the project report was made public by RISCS in October 2017.

The project was led by Thomas Groß. Kovila Coopamootoo was at the University of Derby at the time and responsible for conducting the search phase and the subsequent qualitative coding. Thomas Groß took responsibility for the quantitative analysis. Both authors contributed on writing this report.

This research was supported by the UK Research Institute in the Science of Cyber Security (RISCS), funded through a 2016 National Cyber Security Centre (NCSC) grant on "Scientific Methods in Cyber Security – Systematic Evaluation and Community Knowledge Base for Evidence-Based Methods in Cyber Security."

Thomas Groß was supported by the ERC Starting Grant CASCAde (GA nº716980) for later parts of this investigation.

REFERENCES

[1] Evidence-Based Software Engineering (EBSE), "Guidelines for performing systematic literature reviews in software engineering," Keele University and University of Durham, EBSE Technical Report EBSE-2007-01, July 2007.

[2] K. P. Coopamootoo and T. Groß, "A codebook for experimental research: The nifty nine indicators v1.0," Newcastle University, Tech. Rep. 1514, November 2017.

[3] American Psychological Association (APA), *Publication manual*, 6th ed. American Psychological Association, 2009.

[4] G. Cumming and R. Calin-Jageman, *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge, 2016.

[5] R. Maxion, "Making experiments dependable," *Dependable and Historic Computing*, pp. 344–357, 2011.

[6] S. Peisert and M. Bishop, "How to design computer security experiments," in *Fifth World Conference on Information Security Education*. Springer, 2007, pp. 141–148.

[7] K. P. Coopamootoo and T. Groß, "Evidence-based methods for privacy and identity management," in *Privacy and Identity Management. Facing up to Next Steps*. Springer, 2016, pp. 105–121.

[8] S. Schechter, "Common pitfalls in writing about security and privacy human subjects experiments, and how to avoid them," *Microsoft, January*, 2013.

[9] O. S. Collaboration *et al.*, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, 2015.

[10] B. Everitt, *Cambridge dictionary of statistics*. Cambridge University Press, 1998.

[11] M. B. Brewer, "Research design and issues of validity," *Handbook of research methods in social and personality psychology*, pp. 3–16, 2000.

[12] S. Miller, *Experimental design and statistics*. Routledge, 2005.

[13] R. S. Nickerson, "Null hypothesis significance testing: a review of an old and continuing controversy." *Psychological methods*, vol. 5, no. 2, p. 241, 2000.

[14] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[15] J. Cohen, "A power primer." *Psychological bulletin*, vol. 112, no. 1, p. 155, 1992.

[16] R. E. Kirk, "The importance of effect magnitude," *Handbook of research methods in experimental psychology*, pp. 83–105, 2003.

[17] G. Cumming, "The new statistics: Why and how," *Psychological science*, vol. 25, no. 1, pp. 7–29, 2014.

[18] S. E. Maxwell and H. D. Delaney, *Designing experiments and analyzing data: A model comparison perspective*, 2nd ed. Psychology Press, 2004, vol. 1.

[19] S. Goodman, "A dirty dozen: twelve p-value misconceptions," in *Seminars in hematology*, vol. 45, no. 3. Elsevier, 2008, pp. 135–140.

[20] J. P. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, no. 8, p. e124, 2005.

[21] T. D. Cook and D. T. Campbell, *Quasi-experimentation: Design and analysis for field settings*. Rand McNally, 1979.

[22] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. New York: Psychology Press, Taylor & Francis Group, LCC, 1988.

[23] S. Wacholder, S. Chanock, M. Garcia-Closas, N. Rothman *et al.*, "Assessing the probability that a positive report is false: an approach for molecular epidemiology studies," *Journal of the National Cancer Institute*, vol. 96, no. 6, pp. 434–442, 2004.

[24] S. E. Maxwell, K. Kelley, and J. R. Rausch, "Sample size planning for statistical power and accuracy in parameter estimation," *Annu. Rev. Psychol.*, vol. 59, pp. 537–563, 2008.

[25] D. Moher, A. R. Jadad, G. Nichol, M. Penman, P. Tugwell, and S. Walsh, "Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists," *Controlled clinical trials*, vol. 16, no. 1, pp. 62–73, 1995.

VII. Appendix: SLR Sample

AcqGro2006 - Imagined Communities Awareness Information Sharing and Privacy on Facebook - PETS 2006
AdAcBr2013 - Sleights of Privacy Framing disclosures and the limits of transparency - SOUPS 2013
AfBrGr2012 - Detecting Hoaxes Frauds and Deception in Writing Style Online - IEEE S&P 2012
AfCaSt2014 - Doppelganger Finder Taking Stylometry to the Underground - IEEE S&P 2014
AgShJa2013 - Do not embarass Re-examining user concerns for online tracking and advertising - SOUPS 2013
AhmIss2007 - A New Biometric Technology Based on Mouse Dynamics - IEEE TDSC 2007
AkhPor2013 - Alice in warningland a large-scale field study of browser security warning effectiveness - USENIX 2013
AlbMai2015 - Evaluating the Effectiveness of Using Hints for Autobiographical Authentication A field Study - SOUPS 2015
AlFaWr2015 - The Impact of Cues and User Interaction on the Memorability of System Assigned Recognition-Based Graphical Passwords - SOUPS 2015
AlPoRe2014 - Your Reputation Precedes You History Reputation and the Chrome Malware Warning - SOUPS 2014
AngOrt2015 - WTH Experiences Reactions and Expectations Related to Online Privacy Panic Situations - SOUPS 2015
AtBoHe2015 - Leading Johnny to Water Designing for Usability and Trust - SOUPS 2015
BaMaLi2014 - The Privacy and Security Behaviors of Smartphone App Developers - USEC 2014
BeGiKr2015 - User Acceptance Factors for Anonymous Credentials - WEIS 2015
BeLoSi2007 - Establishing Darknet Connections An evaluation of Usability and Security - SOUPS 2007
BelShe2016 - Crowdsourcing for Context Regarding Privacy in Beacon Encounters via Contextual Integrity - PETS 2016
BenRei2013 - Should users be informed On risk-perception between Android and iPhone users - SOUPS 2013
BeWaLi2010 - The Impact of Social Navigation on Privacy Policy Configuration - SOUPS 2010
BiCoIn2015 - What the App is That Deception and Countermeasures in the Android User Interface - IEEE S&P 2015
BonSch2014 - Towards reliable storage of 56-bit secrets in human memory - USENIX 2014
BoSaRe2012 - Neuroscience Meets Cryptography Designing Crypto Primitives Secure Against Rubber Hose Attacks - USENIX 2012
BrCrDo2013 - Your Attention Please - Designing security-decision UIs to make genuine risks harder to ignore - SOUPS 2013
BrCrKo2014 - Harder to Ignore - Revisiting Pop-up Fatigue and Approaches to Prevent it - SOUPS 2014
BrGrSt2011 - Indirect content privacy surveys - measuring privacy without asking about it - SOUPS 2011
BruVil2007 - Improving Security Decisions with Polymorphic and Audited Dialogs - SOUPS 2007
BrViDj2008 - Evaluating the Usability of Usage Controls in Electronic Collaboration - SOUPS 2008
BuBeFa2010 - How good are Humans at Solving CAPTCHAs - A Large Scale Evaluation - IEEE S&P 2010
BuBePa2011 - The failure of Noise-Based Non-Continuous Audio Captchas - IEEE S&P 2011
BuWoVo2014 - Introducing Precautionary Behavior by Temporal Diversion of Voter Attention from Casting to Verifying their Vote - USEC 2014
CaMiVa2016 - Hidden Voice Commands - USENIX 2016
CaoIve2006 - Intentional Access Management - Making Access Control Usage for End-Users - SOUPS 2006
ChBiOr2007 - A second look at the usability of click-based graphical passwords - SOUPS 2007
ChBoKa2014 - On the Effectiveness of Obfuscation Techniques in Online Social Networks - PETS 2014
ChChBa2015 - You shouldnt collect my secrets - Thwarting sensitive keystroke leakage in mobile IME apps - USENIX 2015
ChMuAs2015 - On the impact of touch id on iphone passcodes - SOUPS 2015
ChObSt2009 - Sanitizations slippery slope- the design and study of a text revision assistant - SOUPS 2009
ChPoSe2012 - Measuring user confidence in smartphone security and privacy - SOUPS 2012
ChStFo2012 - Persuasive cued click-points - Design implementation and evaluation of a knowledge-based authentication mechanism - IEEE TDSC 2012
CzDeYa2010 - Parenting from the pocket - Value tensions and technical directions for secure and private parent-teen mobile safety - SOUPS 2010
DaKrDa2014 - Increasing security sensitivity with social proof - A large-scale experimental confirmation - CCS 2014
DaPuRa2012 - Impact of spam exposure on user engagement - USENIX 2012

DewKul2006 - Aligning usability and security - a usability study of Polaris - SOUPS 2006
DuHeAs2010 - A closer look at recognition-based graphical passwords on mobile devices - SOUPS 2010
DuNiOl2008 - Securing passfaces for description - SOUPS 2008
EgJaPo2014 - Are you ready to lock - CCS 2014
FaFeSh2015 - Anatomization and Protection of Mobile Apps Location Privacy Threats - USENIX 2015
FaHaAc2013 - On the ecological validity of a password study - SOUPS 2013
FaHaMu2012 - Helping Johnny 2.0 to encrypt his Facebook conversations - SOUPS 2012
FoChOo2008 - Improving text passwords through persuasion - SOUPS 2008
GaCaCo2012 - Risk communication design - video vs. text - PETS 2012
GaCaMa2011 - Designing risk communication for older adults - SOUPS 2011
GaChLi2014 - Effective risk communication for android apps - IEEE TDSC 2014
GawFel2006 - Password management strategies for online accounts - SOUPS 2006
GiEgCr2006 - Power Streip Prophylactics and Privacy Oh My - SOUPS 2006
GrCoAl2016 - Effect of cognitive depletion on password choice - LASER 2016
GroBar2014 - Social status and the demand for security and privacy - PETS 2014
HaChDh2008 - Use your illusion- secure authentication usable anywhere - SOUPS 2008
HaChHa2009 - New directions in multisensory authentication - SOUPS 2009
HaCrKl2014 - Targeted threat index - Characterizing and quantifying politically-motivated targeted malware - USENIX 2014
HaDeSm2015 - Where Have You Been - Using Location-Based Security Questions for Fallback Authentication - SOUPS 2015
HaRiSt2012 - Goldilocks and the two mobile devices - going beyond all-or-nothing access to a devices applications - SOUPS 2012
HaScWr2014 - Applying psychometrics to measure user comfort when constructing a strong password - SOUPS 2014
HaZeFi2014 - Its a hard lock life - A field study of smartphone un-locking behavior and risk perception - SOUPS 2014
HuMoWa2012 - Clickjacking - attacks and defenses - USENIX 2012
HuOhKi2015- Surpass - System-initiated user-replaceable passwords - CCS 2015
JaRaBe2014 - To authorize or not authorize - helping users review access policies in organizations - SOUPS 2014
JeSaJe2007 - Tracking website data-collection and privacy practices with the iWatch web crawler - SOUPS 2007
JoEgBe2012 - Facebook and privacy - its complicated - SOUPS 2012
JusAsp2009 - Personal choice and challenge questions - a security and usability assessment - SOUPS 2009
KaBrDa2014 - Privacy Attitudes of Mechanical Turk Workers and the US Public - SOUPS 2014
KaFlRo2010 - Two heads are better than one - security and usability of device associations in group scenarios - SOUPS 2010
KaMaSo2015 - Sound-proof - Usable two-factor authentication based on ambient sound - USENIX 2015
KaTyWa2009 - Conditioned-Safe Ceremonies and a User Study of an Application to Web Authentication - SOUPS 2009
KayTer2010 - Textured agreements - re-envisioning electronic consent - SOUPS 2010
KeBrCr2009 - A nutrition label for privacy - SOUPS 2009
KeCaLi2012 - Self-identified experts lost on the interwebs - The importance of treating all results as learning experiences - LASER 2012
KhHeVo2015 - Usability and security perceptions of implicit authentication - Convenient secure sometimes annoying - SOUPS 2015
KilMax2012 - Free vs. transcribed text for keystroke-dynamics evaluations - LASER 2012
KluZan2009 - Balancing usability and security in a video CAPTCHA - SOUPS 2009
KorBoh2014 - Too Much Choice - End-User Privacy Decisions in the Context of Choice Proliferation - SOUPS 2014
KoShCr2014 - Telepathwords - Preventing weak passwords by reading users minds - SOUPS 2014
KoSoTs2009 - Serial hook-ups - a comparative usability study of secure device pairing methods - SOUPS 2009
KrHuHo2016 - Use the Force- Evaluating Force-Sensitive Authentication for Mobile Devices - SOUPS 2016
KuCrAc2009 - School of phish - a real-world evaluation of anti-phishing training - SOUPS 2009
KuRoCr2006 - Human selection of mnemonic phrase-based passwords - SOUPS 2006
LeMoPe2016 - Privacy Challenges in the Quantified Self Movement - An EU Perspective - PETS 2016

LiAnSc2016 - Follow my recommendations - A personalized privacy assistant for mobile app permissions - SOUPS 2016
LiAsCa2008 - Risk communication in security using mental models - USEC 2008
LiBrYe2011 - Demographic Profiling from MMOG Gameplay - PETS 2011
LiLiSa2014 - Modeling users' mobile app privacy preferences - Restoring usability in a sea of permission settings - SOUPS 2014
LiXiPe2011 - Smartening the crowds- computational techniques for improving human verification to fight phishing scams - SOUPS 2011
LlPoAt2015 - Face-off - Preventing Privacy Leakage From Photos in Social Networks - CCS 2015
MaDeKe2011 - Using data type based security alert dialogs to raise online security awareness - SOUPS 2011
MaLeAd2012 - The PViz comprehension tool for social network privacy settings - SOUPS 2012
MalPre2013 - Sign-up or give-up- Exploring user drop-out in web service registration - SOUPS 2013
MoGaSa2014 - Dynamic cognitive game captcha usability and detection of streaming-based farming - USEC 2014
MohaBe2010 - Do windows users follow the principle of least privilege - investigating user account control practices - SOUPS 2010
MoLiVi2014 - Understanding and specifying social access control lists - SOUPS 2014
NoBlCa2014 - Why Johnny Cant Blow the Whistle - Identifying and Reducing Usability Issues in Anonymity Systems - USEC 2014
PanCut2010 - Usably secure low-cost authentication for mobile banking - SOUPS 2010
PaNoKa2012 - Reasons rewards regrets - privacy considerations in location sharing as an interactive practice - SOUPS 2012
PanPra2014 - Crowdsourcing attacks on biometric systems - SOUPS 2014
PeKoBu2014 - Cloak and swagger - Understanding data sensitivity through the lens of user anonymity - IEEE S&P 2014
PoHaEg2012 - Android permissions - User attention comprehension and behavior - SOUPS 2012
PoIlMa2014 - Faces in the distorting mirror- Revisiting photo-based social authentication - CCS 2014
PuGros2015 - Towards a Model on the Factors Influencing Social App Users Valuation of Interdependent Privacy - PETS 2015
RaBoJa2014 - To befriend or not - a model of friend request acceptance on facebook - SOUPS 2014
RaDeGr2016 - Privacy Wedges- Area-Based Audience Selection for Social Network Posts - SOUPS 2016
Rader2014 - Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google - SOUPS 2014
RaHaBe2009 - Revealing hidden context- improving mental models of personal firewall users - SOUPS 2009
RajCam2016 - Influence of Privacy Attitude and Privacy Cue Framing on Android App Choices - SOUPS 2016
RaWaBr2012 - Stories as informal lessons about security - SOUPS 2012
ReKrMa2016 - How I Learned to be Secure- a Census-Representative Survey of Security Advice Sources and Behavior - CCS2016
RiBoMo2016 - Measuring the influence of perceived cybercrime risk on online service avoidance - IEEE TDSC 2016
RiQiSt2012 - Progressive authentication- deciding when to authenticate on mobile phones - USENIX 2012
RoCuJo2014 - Behavioral Experiments Exploring Victims Response to Cyber-based Financial Fraud and Identity Theft Scenario Simulations - SOUPS 2014
RuKiBu2013 - Confused Johnny- when automatic encryption leads to confusion and mistakes - SOUPS 2013
RuOnYo2016 - User Attitudes Toward the Inspection of Encrypted Traffic - SOUPS 2016
SchBon2015 - Learning assigned secrets for unlocking mobile devices - SOUPS 2015
SchRee2009 - 1 plus 1 equal you- measuring the comprehensibility of metaphors for configuring backup authentication - SOUPS 2009
ScMcPa2011 - Empowering end users to confine their own applications - The results of a usability study comparing SELinux AppArmor and FBAC-LSM - TISSEC 2011
ScWaKo2013 - Exploring the design space of graphical passwords on smartphones - SOUPS 2013
ShBeRo2016 - Behavioral Study of Users When Interacting with Active Honeytokens - TISSEC 2016
ShKeKo2012 - Correct horse battery staple- Exploring the usability of system-assigned passphrases - SOUPS 2012
ShKoDu2016 - Designing Password Policies for Strength and Usability - TISSEC 2016
ShKoKe2010 - Encountering stronger password requirements- user attitudes and behaviors - SOUPS 2010
ShKrVi2015 - Portrait of a Privacy Invasion - PETS 2015

ShKuSe2014 - Beware your hands reveal your secrets - CCS 2014
ShMaKo2007 - Anti-phishing phil- the design and evaluation of a game that teaches people not to fall for phish - SOUPS 2007
SmeGoo2009 - How users use access control - SOUPS 2009
StHuBr2012 - Are privacy concerns a turn-off- engagement and privacy in social networks - SOUPS 2012
StoBid2013 - Memory retrieval and graphical passwords - SOUPS 2013
SuEgAl2009 - Crying Wolf - An Empirical Study of SSL Warning Effectiveness - USENIX 2009
TaOzHo2006 - A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords - SOUPS 2006
ThLiCh2016 - What Questions Remain - An Examination of How Developers Understand an Interactive Static Analysis Tool - SOUPS 2016
UrKeKo2012 - How does your password measure up - the effect of strength meters on password creation - USENIX 2012
VItak2015 - Balancing privacy concerns and impression management strategies on Facebook - SOUPS 2015
WaGeCh2016 - On the Security and Usability of Segment-based Visual Cryptographic Authentication Protocols - CCS 2016
WaRaBe2016 - Understanding Password Choices - How Frequently Entered Passwords are Re-used Across Websites - SOUPS 2016
WrPaBi2012 - Do you see your password- applying recognition to textual passwords - SOUPS 2012
WuMiLi2006 - Web wallet- preventing phishing attacks by revealing user intentions - SOUPS 2006
XuReCh2012 - Security and usability challenges of moving-object CAPTCHAs- decoding codewords in motion - USENIX 2012
YaLiCh2016 - An Empirical Study of Mnemonic Sentence-based Password Generation Strategies - CCS 2016
YeHeOp2014 - An epidemiological study of malware encounters in a large enterprise - CCS 2014
ZhPaWa2016 - An Efficient User Verification System Using Angle-Based Mouse Movement Biometrics - TISSEC 2016
ZhWaJi2014 - Privacy Concerns in Online Recommender Systems- Influences of Control and User Data Input - SOUPS 2014